Available online at www.sciencedirect.com
ScienceDirect
Journal of Physiology - Paris 101 (2007) 110–117
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

# Estimating the hidden learning representations

Andrea Brovelli [a,*], Pierre-Arnaud Coquelin [b], Driss Boussaoud [a]

[a] Institut de Neurosciences Cognitives de la Méditerrannée, UMR 6193 CNRS-Université de la Méditerranée, 31 chemin Joseph Aiguier, 13402 Marseille, France
[b] Centre de Mathématiques Appliquées, UMR 7641 CNRS-Ecole polytechnique, 91128 Palaiseau, France

**Abstract**

Successful adaptation relies on the ability to learn the consequence of our actions in different environments. However, understanding the neural bases of this ability still represents one of the great challenges of system neuroscience. In fact, the neuronal plasticity changes occurring during learning cannot be fully controlled experimentally and their evolution is hidden. Our approach is to provide hypotheses about the structure and dynamics of the hidden plasticity changes using behavioral learning theory. In fact, behavioral models of animal learning provide testable predictions about the hidden learning representations by formalizing their relation with the observables of the experiment (stimuli, actions and outcomes). Thus, we can understand whether and how the predicted learning processes are represented at the neural level by estimating their evolution and correlating them with neural data. Here, we present a bayesian model approach to estimate the evolution of the internal learning representations from the observations of the experiment (state estimation), and to identify the set of models' parameters (parameter estimation) and the class of behavioral model (model selection) that are most likely to have generated a given sequence of actions and outcomes. More precisely, we use Sequential Monte Carlo methods for state estimation and the maximum likelihood principle (MLP) for model selection and parameter estimation. We show that the method recovers simulated trajectories of learning sessions on a single-trial basis and provides predictions about the activity of different categories of neurons that should participate in the learning process. By correlating the estimated evolutions of the learning variables, we will be able to test the validity of different models of instrumental learning and possibly identify the neural bases of learning.


*Keywords:* Arbitrary visuomotor learning; Bayesian model; Sequential Monte Carlo methods; Maximum likelihood principle

## 1. Introduction

Learning and remembering the consequences of our actions is one of the most fundamental forms of intelligent behavior, because it allows us, as well as other animals, to anticipate relevant events and adapt to changing environments. This ability can be studied within the framework of instrumental learning, where an animal learns by trial-and-error the contingency arranged by a given stimulus, or context, between an action and its outcome (Rescorla, 1991; Dickinson, 1994).

In monkey neurophysiology, research on the neural bases of instrumental learning has relied on two complementary approaches. One searched for co-variations of the firing rate of single neurons with the improvement of performance, measured by the probability of correct response. Using this approach, three classes of neurons, whose activity changes correlate positively or negatively with the probability of correct response or with its rate of change, have been found in the dorsal premotor and prefrontal cortex, striatum and hippocampus (Wise and Murray, 2000; Brasted and Wise, 2005; Suzuki and Brown, 2005). A complementary approach searched for changes in the selectivity of neuronal activity for the rewarded motor response in the average firing rate of neuronal populations (Asaad et al., 1998; Pasupathy and Miller, 2005).

* Corresponding author.
*E-mail address:* andrea.brovelli@incm.cnrs-mrs.fr (A. Brovelli).

0928-4257/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2007.10.002

These authors found that the selectivity strength of neurons in the prefrontal cortex and in the striatum increases and its latency decreases (occurs earlier in the trial) as learning takes place.

Overall, these studies showed how the neural activity in the frontal cortex, basal ganglia and hippocampus evolves during arbitrary visuomotor learning. However, it is not clear which hidden learning representations (e.g., the subjective "value" of a given action) are coded by these learning-related neural activities. Our approach is to provide hypotheses about the structure and dynamics of the hidden learning representations through behavioral learning models. In fact, behavioral models of animal learning provide testable predictions about the hidden learning representations by formalizing their relation with the observables of the experiment (stimuli, actions and outcomes). Thereafter, we can understand whether and how the predicted learning processes are represented at the neural level by estimating their evolution and correlating them with neural data.

This paper focuses on the first part of this approach. More precisely, our objective is to provide a general mathematical method to estimate the evolution of the internal learning representations from the observations of the experiment (state estimation), together with the set of models' parameters (parameter estimation) and the class of behavioral model (model selection) that are most likely to have generated a given sequence of actions and outcomes. To do so, we use a bayesian methodology for state estimation and the maximum likelihood principle (MLP) for model selection and parameter estimation. To estimate the posterior probabilities arising in the estimation of the state and in the evaluation of the likelihood, we used sequential Monte Carlo methods (SMC) (for a review see Doucet and Godsill, 2000; Doucet et al., 2001), because they are able to cope with learning models containing non-linear terms and non-Gaussian distributions. We tested this method on simulated sequences of behavioral events according to two of the most plausible models of instrumental learning, and studied its accuracy in model selection, parameter and state estimation. The models provide diverging predictions about the evolution of the internal representations of the learning process. These predictions qualitatively fit the electrophysiological results available in the literature and provide a theoretical framework for interpreting the changes occurring at neural level during learning.

## 2. Methods

### 2.1. Behavioral models of instrumental learning

Since our goal is to provide a method for model selection, parameter and hidden state estimation, we considered two behavioral models. Each model provides a different account of the internal learning processes, through specific learning rules and free parameters. The first is based on animal associative learning theory (Dickinson, 1997; Pearce and Bouton, 2001; Schultz and Dickinson, 2000), whereas the second is a non-associative model based on the Matching law (Herrnstein, 1961; Herrnstein, 1970) and previously used to model discrete-choice decision-making tasks (Sugrue et al., 2004; Sugrue et al., 2005). Our implementation also contains formalisms classically used in reinforcement learning algorithms (Sutton and Barto, 1998).

#### 2.1.1. The associative model

Associative learning theory postulates that the ability to learn the consequence of a particular behavior in a given environment resides in the formation of stimulus–response–outcome associations whose strength varies according to the contingency and contiguity of the events, as well as the current goals and motivational state of the animal (Rescorla, 1991; Dickinson, 1994; Balleine and Dickinson, 1998). If only one stimulus $s$ is needed to set the contingency (as in the following simulations), the behavioral change (e.g., the probability to perform a given joystick movement to obtain reward) is assumed to reflect a gradual strengthening of the internal representation of the action–outcome association brought about by each pairing (Dickinson, 1997). This means that on a given trial during learning, only the associative value for the chosen action is updated. If more than stimulus is present in the learning session and if each sets specific response–outcome contingencies, only the associative value for the presented stimulus and chosen action is updated. Since the associative model does not formalize the interactions between separate stimulus–response–outcome associations, we will simulate learning sessions when only one stimulus $s$ is needed to set the response–outcome contingencies. The following mathematical description in Section 2.2 will however let the stimuli vary over time, for generality reasons.

In order to quantify the evolution of the associative values, several associative models have been developed (Rescorla and Wagner, 1972; Macintosh, 1975; Pearce and Hall, 1980). One of the most influential learning model is the one developed by Rescorla and Wagner Rescorla and Wagner (1972), where the evolution of the associative strengths for each action is given by

$$V^a_{t+1} = V^a_t + \eta(r(t) - V^a_t) + \epsilon, \tag{1}$$

where $t$ is trial number, $a \in \{1, \ldots, n\}$ is the action (where $n$ is the number of possible actions), $\eta$ is the learning rate for a given stimulus and $\epsilon$ is gaussian noise $N(0, \sigma)$. The asymptotic value of the association strength $r(t)$ is 1 for a correct response, 0 if incorrect. Thus, the change in associative strength on a particular trial (i.e., the amount of learning on each trial) is equal to the learning rate $\eta$ times the error term $(r(t) - V^a_t)$, which is referred to as the prediction-error signal (Schultz, 2006). This value is the difference between the maximum associative strength and the current prediction of the associative strength. In other words, it represents the discrepancy between the obtained and expected outcome. The subjective representation of the expected outcome prior to learning is formalized in the associative model as the initial value of the associative strengths $V^a_{t=1}$. These values can be set to zero, meaning that the subjects do not have prior expectancies about a correct outcome after a given movement. However, since the number of possible actions is $n$ and given that only one of the actions was considered as correct in the following simulations, we set the prior subjective probability of correct response to $1/n$.

Associative learning theory does not state how the associative strengths are "translated" into behavior. However, we expect some degree of stochasticity in the action selection process, where the probability to perform a given action is proportional to its associative value. To model such an action selection process, we transformed the association values $V^a_t$ into probabilities according to the softmax equation, which is a standard method in reinforcement learning theory (Sutton and Barto, 1998)

$$\mathbb{P}(a_t, V^a_t) \triangleq \frac{e^{\beta V^a_t}}{\sum_a e^{\beta V^a_t}}. \tag{2}$$

The coefficient $\beta$ is the inverse "temperature": low $\beta$ values cause the actions to be all (nearly) equiprobable, whereas high $\beta$ amplify the differences in association values. As $\beta$ tends to infinity the softmax equation reduces to a winner-take-all function assigning a probability equal 1 to the action with the highest association value. Therefore, we can modulate the degree of stochasticity in action selection by varying the value of this

model parameter. In reinforcement learning theory (Sutton and Barto, 1998), this model is also referred to as the Q-learning algorithm (Watkins and Dayan, 1992), in which action values are updated through the Rescorla–Wagner learning rule. Overall, the internal learning representations (state process) hypothesized by the associative model are the associative values and the prediction-error signals.

#### 2.1.2. The total-income model

Instrumental learning can also be described according to a non-associative model based on the operant matching law, which states that an organism allocates its behavior over various activities in exact proportion to the value derived from each activity (Herrnstein, 1961; Herrnstein, 1970). In fact, the matching law can also be applied to produce a simple learning model (Seung, 2004), where an animal selects actions to match the total rewards earned (total income) from each action in discrete-choice decision-making tasks (Sugrue et al., 2004; Sugrue et al., 2005). Since the action selection process is determined by the reward history, we modeled the evolution of the total income $I_t^a$ during learning as

$$I_{t+1}^a = I_t^a + \eta r(t) + \epsilon, \tag{3}$$

where $a \in \{1,\ldots,n\}$ is the action (where $n$ is the number of possible actions), $\eta$ determines the slope of the increase (learning rate). The value of the reward $r(t)$ is equal to 1 when the reinforcement is present and 0 when the reinforcement is absent. The total income prior to learning is zero and it is formalized in the present model as the initial values of total income $I_{t=1}^a$. During learning, the total income $I_t^a$ for the rewarded action approximates a linearly increasing function of the number of rewards obtained. The action selection process computes the probability to perform the actions proportionally to the income values as in the previous model according to Eq. (2).

### 2.2. A mathematical methodology for estimation in behavioral sequential learning tasks

#### 2.2.1. Mathematical structure of sequential learning experiments

We suppose that we can observe a sequence of stimulus–action-reward from time 1 to time $T$: $(s_t,a_t,r_t)_{t\in\{1\ldots,T\}}$. Note that the following simulations contain one 1 stimulus; here we present the generalized situation when more than one stimulus is present by letting $s_t$ vary over time. We want to recover the sequence of the subjective learning state $(X_t)_{t\in\{1\ldots,T\}}$ that has been used by the learner to act during this experiment. For example, in the case of the Rescola–Wagner model $X_t = V_t$ (see Eq. (1)) and in the case of the total income model $X_t = I_t$ (see Eq. (3)). The learning rule (e.g., Eq. (1) or Eq. (3)) and the action rule (e.g., Eq. (2)) provide a full probabilistic structure to the learning experiment which is represented in Fig. 1.

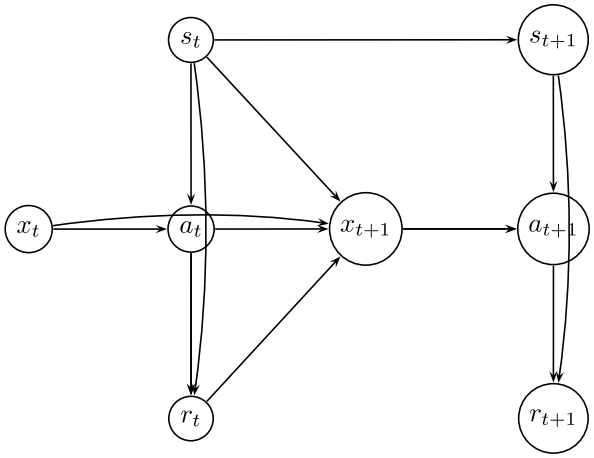


Fig. 1. Probabilistic structure of the learning process. $s_t$ is the stimulus presented at time $t$, $a_t$ is the action made at time $t$, $r_t$ is the reward received at time $t$ and $x_t$ is the internal learning state of the animal at time $t$.

#### 2.2.2. Estimation of the (hidden) internal learning state

Under the probabilistic structure depicted in Fig. 1, the knowledge of the (hidden) value of internal learning state at time $t$ is summarized in the posterior probability distribution $\pi_t$ of the learning state value $X_t$ given the observations up to time $t$

$$\pi_t(X_t^a) \triangleq \mathbb{P}(X_t^a|(s_t,a_t,r_t)_{t\in\{1\ldots,T\}}).$$

Note that $X_t$ is a vector whose coordinates are $X_t^a$ for $a \in \{1,\ldots,n\}$. Estimating $\pi_t$ analytically is in general impossible. However, Sequential Monte Carlo methods (SMC) (see Doucet et al., 2001; Doucet and Godsill, 2000 for a good introduction to SMC) approximate $\pi_t$ by an empirical distribution of "particles"

$$\pi_t^N \triangleq \frac{1}{N}\sum_i \delta_{x_t^i},$$

where the particles $x_t^i$ are obtained using Algorithm 1, and $\delta_x$ denote the Dirac mass at the point $x$. From this approximation one can deduce a pointwise approximation of $X_t$ by evaluating its mean

$$\hat{x}_t = \frac{1}{N}\sum_i x_t^i.$$

One can also estimate confidence interval by evaluating its covariance matrix, for $1 \leqslant a, a' \leqslant n$ define

$$\sigma_t^{aa'} = \frac{1}{N^2}\sum_i (x_t^{i,a} - \hat{x}_t^a)(x_t^{i,a'} - \hat{x}_t^{a'}).$$

Algorithm 1. Interacting Particles for Associative Learning (IPAL)

**for t = 1 to *T* do**
- For all $i \in \{1,\ldots,N\}$, sample independently $\tilde{x}_t^i$ from $x_{t-1}^i$ using the learning equation (e.g. Eq. (1) or Eq. (3))
- For all $i \in \{1,\ldots,N\}$, associate to the $i$th particle a weight $w_t^i$ proportional to the probability of doing the action $a_t$ in the internal learning state $x_t^i$ normalized with respect to the sum over particles: $w_t^i \triangleq \frac{P^{a_t}_{x_t^i}}{\sum_{j=1}^N P^{a_t}_{x_t^j}}$
- For all $i \in \{1,\ldots,N\}$, sample the $i$-selection index $j_i$ from the multinomial distribution defined as $\mathbb{P}(j_i = j) = w_t^j$
- For all $i \in \{1,\ldots,N\}$, select the particles using $x_t^i \triangleq \tilde{x}_t^{j_i}$

**end for**

#### 2.2.3. Model selection and parameter estimation

To achieve model selection and parameter estimation we propose to use the maximum likelihood principle (MLP) (see Myung, 2003 for a tutorial about the use of MLP). Indeed, denoting by $\theta$ the vector of parameters associated to the learning model (e.g., $\theta = (\eta,\beta)$), the log-likelihood function evaluated at the parameter value $\theta$ is defined by

$$l(\theta) \triangleq \log \mathbb{P}_\theta((s_t,a_t,r_t)_{t\in\{1\ldots,T\}}). \tag{4}$$

The maximum likelihood principle consists in estimating the unknown parameter $\theta^*$ used by the learner by maximizing the log-likelihood

$$\hat{\theta}^* \triangleq \operatorname*{argmax}_{\theta\in\Theta} l(\theta),$$

where $\Theta$ is the domain of possible parameter value and it depends on the specific learning model that is used. The log-likelihood using a Sequential Monte Carlo methods is given by Doucet and Godsill (2000)

$$l(\theta) = \sum_{t=1}^T \log\left(\frac{1}{N}\sum_{i=1}^N w_t^i\right),$$

where $w_t^i$ is obtained by applying Algorithm 1 with the parameter value $\theta$. To maximize the log-likelihood one can use any usual stochastic optimization algorithm (e.g. annealed simulated, genetic algorithms, Nelder-Mead simplex, grid based approach, etc.; see Spall, 2003 for a review). In the case of large dimensional parameter space, on can use a gradient-based method

(Coquelin et al., 2007). For model selection, we compute the optimal parameter $\theta_k^*$ for each model $k$, and the more accurate model $k^*$ is the one with the highest log-likelihood at the optimal parameter value. In the present paper we simply used 2 models (i.e., $k = 2$) and a grid based approach described in the next section.

### 2.3. Numerical simulations

We simulated the behavioral choices of an animal learning by trial-and-error the correct association between an action (e.g., joystick movements) and its outcome (e.g., reward). Three actions are possible and only one is rewarded. We generated the behavioral choices according to either the associative or the total-income model. Since, for a given model, the evolution of the learning variables (e.g., the associative strengths) depends on the values of the models' parameters $\eta$ and $\beta$, we varied $\eta = 0.1 \rightarrow 0.3$ (step $= 0.1$) and $\beta = 1 \rightarrow 3$ (step $= 1$). The outcome value $r(t)$ was either 1 or 0 for correct and incorrect actions. The gaussian noise $\epsilon$ added to the state process was drawn from a normal distribution $N(0, \sigma)$, where $\sigma = 0.005$. For a given behavioral model and parameter set $\theta(\eta, \beta)$, we generated 100 learning sessions, each lasting 150 trials. Even though the ranges of parameters' values do not need to be identical for the two models, we chose them identical so to produce similar learning curves, defined as the probability of correct response. More precisely, we wished to simulate learning sessions that reached the asymptotic probability of correct response at approximately the same trial both for the associative and the total-income model. This effect can be seen in Fig. 6 (left panel): the probability of correct response (averaged over parameter and sessions) for the associative (Fig. 6A) and total-income (Fig. 6B) model reaches the asymptotic value at approximately the same trial number for both models. We tested the accuracy of the proposed methodology in model selection, parameter identification and state estimation for each learning session.

## 3. Results

### 3.1. Model selection

We first quantified the ability of the present methodology to correctly identify which model generated a given sequence of observations. Given 100 action–outcome sequences generated using a given model with fixed parameters, we computed the number of sessions where the method successfully identified the true model, by comparing the log-likelihood computed from both models. The selected model was the one corresponding to the highest value of the log-likelihood. The results showed that for all sessions and parameter's sets, the method always selected the true model as the generator.

### 3.2. Parameter identification

We then analyzed the accuracy in identifying the correct values of models' parameters $\eta$ and $\beta$. Fig. 2 shows the joint probability of estimating the true values of both $\eta$ and $\beta$ for the associative model. Each panel corresponds to the joint probabilities computed on 100 sessions generated using a fixed set of parameters, indicated by the grey shade. For example, the top left panel shows the joint probabilities for sessions generated using $\eta = 0.1$ and $\beta = 3$; the correct estimation of both $\eta$ and $\beta$ occurred in 68 sessions out of 100 (grey shade). The probability of correctly identifying $\eta$ independently of $\beta$ can be computed by summing

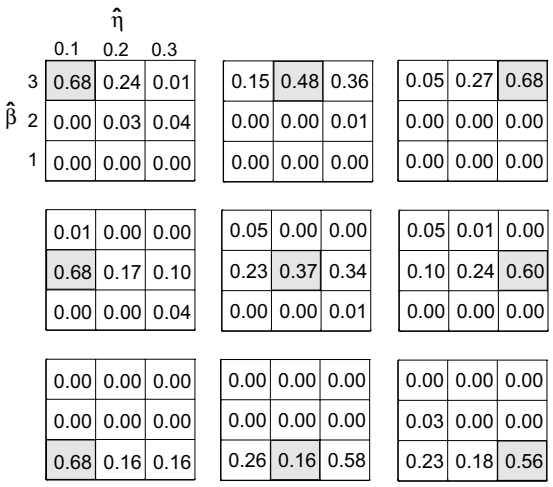


Fig. 2. Joint probability distribution $p(\hat{\eta}, \hat{\beta})$, where $\hat{\eta}$ and $\hat{\beta}$ are the estimated learning rate and inverse "temperature", respectively. Each panel shows the joint probabilities for a given set of true model's parameters, which is indicated by the gray share. For example, the probabilities in the top-left panel were calculated from 100 learning sessions generated using $\eta = 0.1$ and $\beta = 3$; the probability of correctly estimating both model parameters was 0.68.

the joint probabilities horizontally, and vertically for correct $\beta$ estimation. Overall, the results show that the true $\beta$ was more probable to be correctly estimated than the true $\eta$. The largest error in the parameter's estimation occured when the learning rate was equal to 0.2. This phenomenon is more evident when the stochasticity in the associative model increases, that is when the $\beta$ values are equal to 1. To understand whether this is due to the limited range of $\eta$ values (0.1, 0.2 and 0.3) or by a systematic estimation problem for the associative model, we simulated learning sessions with a fixed value for $\beta = 1$ and we varied $\eta$ between 0.05 and 0.4 in steps of 0.05. The results showed that the parameter's estimation is more accurate for low (from 0.05 to 0.15) and high (from 0.3 to 0.4) values of the learning rate, whereas the error reaches a maximum when $\eta$ takes intermediate values. Finally, we analyzed the accuracy in identifying the correct values of models' parameters $\eta$ and $\beta$ for the total-income model (full data not shown). The most reliable estimation (94% of the sessions) occurred for $\eta = 0.1$ and $\beta = 1$. Overall, the simulations allowed us to determine the probability of correct parameter identification as a function of parameters' values. These results have to be taken into account when analyzing real behavioral data.

### 3.3. Dependence of model and parameter estimation on noise level

To quantify the dependence of model selection and parameter estimation on the noise level $\epsilon$, we simulated 100 learning sessions using the associative model with fixed parameters ($\eta = 0.1$ and $\beta = 3$), and we varied the standard deviation of the noise distribution $\sigma$ ($N(0, \sigma)$) from 0.005 to 0.25 in steps of 0.005. The results are shown in Fig. 3: even

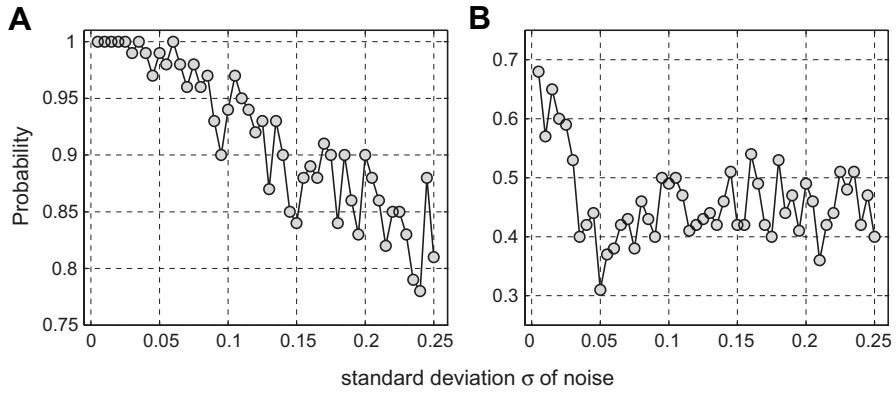


Fig. 3. Dependence of noise level $\epsilon$ on model and parameter estimation. (A) Probability of correct model estimation as a function of $\sigma$; (B) probability of correct parameter estimation as a function of $\sigma$. The noise values $\epsilon$ are drawn from a normal distribution $N(0,\sigma)$.

though the model selection is relatively accurate (above 90%) for $\sigma < 0.125$ (Fig. 3A), the parameter selection quickly drops to 40% for values of $\sigma > 0.025$ (Fig. 3B). These results are exemplar, because the degrading effect due to noise depends on the choice of models parameters: for example, if the stochasticity in the action selection process is high (e.g., $\beta = 1$), the effect of noise on model and parameter estimation will be stronger than for less stochastic models (e.g., $\beta = 3$). However, our simulations allow us to estimate the effect of noise level on model selection and parameter estimation.

### 3.4. State estimation

We then studied how the proposed method performs in the state estimation. The Sequential Monte Carlo method estimates the true learning representation (e.g., the associative strengths) only if the true parameters are identified. Fig. 4A shows the evolution of the simulated and estimated associative values for a representative learning session, in which the correct parameters were estimated. Fig. 4B shows a session where the true parameters were not identified; the distance between the simulated and estimated curves is stronger early during learning, where the variance in the state process is highest. Fig. 5 shows the evolution of the simulated and estimated total-income values for a representative learning session; the difference between the simulated and estimated curves does not attenuate during learning if the model's parameters are not estimated correctly (Fig. 5B). On the other hand, the present method qualitatively retrieves the true learning representations on a single-trial basis, even though the correct parameters are not exactly identified.

Finally, in order to quantify the mean evolution of the learning representations for the two models, we averaged the evolution of the state processes across all parameter sets and learning sessions. Fig. 6A shows the probability of correct response (first panel from the left), the mean evolution of the associative strengths for the performed action (second panel), the mean evolution of the associative strengths for the unplayed action (third panel), and the prediction-error for the performed action (fourth panel). Fig. 6B, shows the corresponding state processes for the total-income model. Therefore, the two models provide diverging predictions about the evolution of the learning representations.These curves provide predictions about

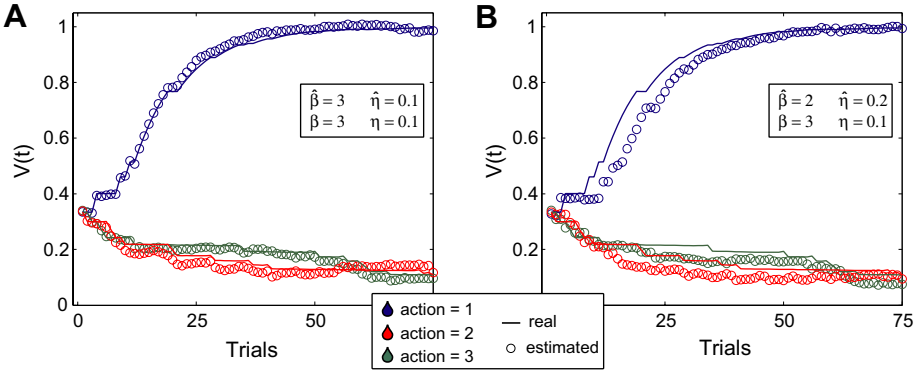


Fig. 4. Single-trial evolution of the hidden learning representation in exemplar sessions. (A) Simulated (thick line) and estimated (circles) learning representations for the associative model in an exemplar session, where the correct model's parameters had been identified. (B) Same as in (A), but for an exemplar session where the correct parameters were not identified by the method. The true and estimated parameters are indicated in the top-right panel in each plot.

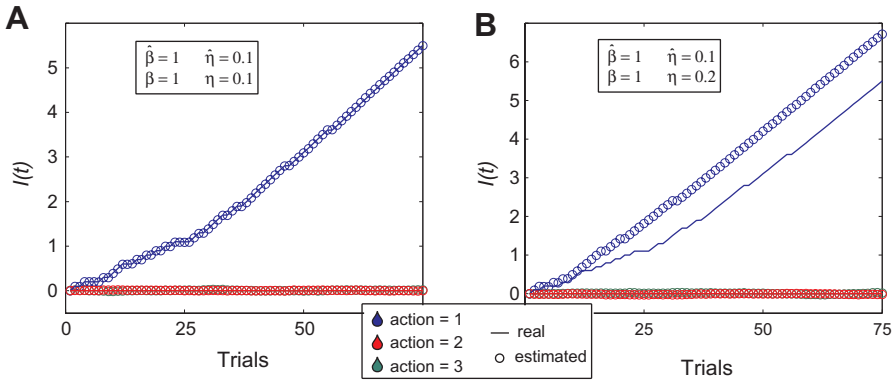


Fig. 5. Single-trial evolution of the hidden learning representation in exemplar sessions. (A) Simulated (thick line) and estimated (circles) learning representations for the total-income model in an exemplar session, where the correct model's parameters had been identified. (B) Same as in (A), but for an exemplar session where the correct parameters were not identified by the method. The true and estimated parameters are indicated in the top panel in each plot.

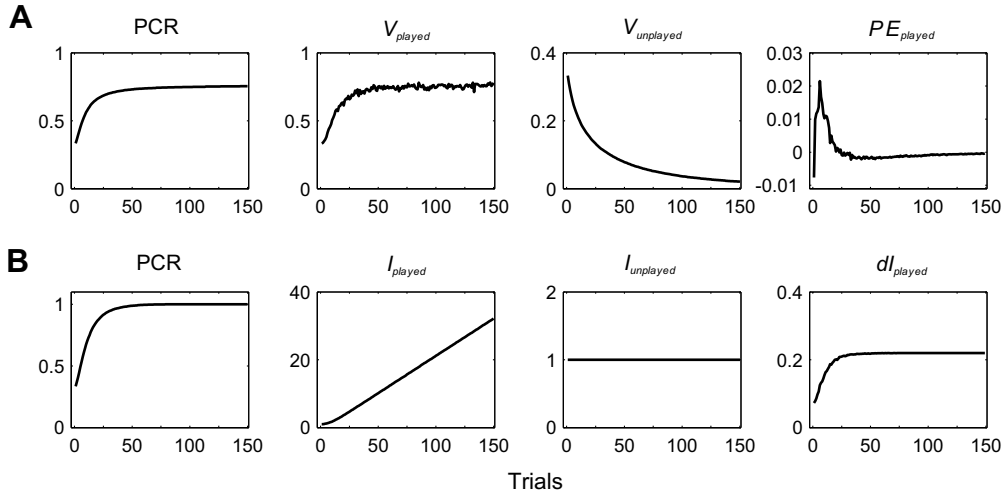


Fig. 6. Average evolution of the learning representations for the associative model (A) and total-income model (B). PCR is the probability of correct response; $V_{\text{played}}$ and $V_{\text{unplayed}}$ are the mean associative values for the played and unplayed actions, respectively; $\text{PE}_{\text{played}}$ is the prediction-error signal for the played action; $I$ is the total-income; $\text{d}I_{\text{played}}$ is the rate of change for the total income of the played action.

the average evolution of the learning representations according to either models; we will discuss the relevance of these results in the next section.

## 4. Discussion

In the present paper, we described and tested a method based on a Bayesian approach to estimate the evolution of the internal learning representations from the observations of the experiment (state estimation), to identify the set of models' parameters (parameter estimation) and the class of behavioral model (model selection) that are most likely to have generated a given sequence of actions and outcomes. More precisely, we used sequential Monte Carlo methods for state estimation and the maximum likelihood principle (MLP) for model selection and parameter estimation. Our simulation study allowed us to quantify the accuracy and ability in model selection, parameter estimation and state identification. The model selection was errorless, meaning that the method identified the correct model as the true generator of a behavioral sequence of actions and outcomes. The reliability depends on how similar the considered models are; in the present case, the evolution of the learning representations according to the associative model and total-income model diverge rapidly during learning. That is why the method accurately identified the true model in all sessions. For what concerns the parameters' estimation, we quantified the range of $\eta$ and $\beta$ for which the true parameters were identified and computed the probability of identifying both $\eta$ and $\beta$ (Fig. 2). This information is helpful for studies analyzing the behavioral responses measured during electrophysiological and/or neuroimaging studies, because it can be used to provide error bounds on estimated model's parameters. Even though the method does not retrieve the true model's parameters in all learning sessions, the estimation of the learning representations (state

processes) qualitatively fits the simulated trajectories (e.g., Figs. 3 and 4). This is crucial if we want to understand whether a given behavioral model provides an accurate description of the neural plasticity changes mediating learning.

Sequential Monte Carlo methods have been previously used to estimate the hidden value of the association strengths (Samejima et al., 2004). However, the authors did not consider the problem of model selection, and tested the accuracy of the method on a single behavioral model. In addition, since the authors considered $\eta$ and $\beta$ as time-varying hyperparameters, their initial estimates were inevitable inaccurate, and attained good estimates after about 200 observations. This produced large deviations from the simulated data especially in the first 50 trials of the learning sessions. In the present paper, we showed that accurate estimation of the state process can be attained on a shorter time scale and we stressed the importance of model selection as a fundamental step in analyzing behavioral data.

The two models provide diverging predictions about the internal processes of learning. The associative Rescorla–Wagner model predicts the existence of a neural substrate coding for the fast increase and decrease in associative strengths and for prediction-error signals (Fig. 5A). The total-income model predicts neurons coding for the quasi-linear increase of the total income (Fig. 5B). We will here review the literature about the neural correlates of conditional visuomotor arbitrary learning and show how our results can provide a better interpretation of the electrophysiological results.

A first set of electrophysiological results suggests that the evolution of the associative strengths could be coded by modulations in firing rate of single neurons. In fact, two classes of neurons showing either a monotonic increase or decrease in firing rate that correlates (either positively or negatively) with the learning curve (i.e., the probability of correct response curve) have been found in the hippocampus (Cahusac et al., 1993; Wirth et al., 2003), striatum (Tremblay et al., 1998; Hadj-Bouziane and Boussaoud, 2003; Brasted and Wise, 2004; Williams and Eskandar, 2006), frontal and eye field (Chen and Wise, 1995a; Chen and Wise, 1995b), dorsal premotor cortex Brasted and Wise, 2004 and orbitofrontal cortex (Tremblay and Schultz, 2000). Since the curve representing the probability of correct response is correlated with the evolution of the association values (Fig. 4A, first and second panel), these two classes of neurons could code for the formation and dissolution of associations. A third category of neurons displaying a modulation in firing rate during learning have been found in the hippocampus (Cahusac et al., 1993), striatum (Hadj-Bouziane and Boussaoud, 2003; Williams and Eskandar, 2006) frontal and supplementary eye field (Chen and Wise, 1995b) and dorsal premotor cortex (Brasted and Wise, 2004). The changes in their firing rate are characterized by an initial increase followed by a decrease to virtually inactivity, with the maximal discharge around the time of learning (i.e., when the rate of change of learning is highest); in addition, these neurons do not typically discharge during the execution of well-known associations. A recent study described a population of striatal neurons whose activity during the feedback sound and reward periods correlates with the rate of learning (estimated from the first of correct response) (Williams and Eskandar, 2006), which resembles the prediction-error curves of the present study (Fig. 5A, fourth study, fourth panel). Therefore, our results together with those present in the literature suggest this third class of neurons might code for prediction-error signals at the cortical level, probably under the influence of neurons from the dopaminergic system (for reviews Schultz, 2006; Schultz and Dickinson, 2000). In other words, we suggest that this third class of neurons is the local responsible for the changes in the activity observed in the first two classes of neurons.

The neural correlates of the total-income model as applied to arbitrary visuomotor learning have not been studied extensively. However, there exists electrophysiological evidence suggesting that the total income accumulating with accumulating rewards could be coded by populations of neurons of the prefrontal cortex. For example, a recent study from (Pasupathy and Miller, 2005), originating from an earlier study by the same group (Asaad et al., 1998), showed that the average strength of direction selectivity during a peri-saccade epoch in a population of prefrontal neurons undergoes a linear increase as a function of correct trials (Pasupathy and Miller, 2005). This evolution cannot be accounted for by the associative model, because the increase in associative strength always follows a negatively accelerating curve. However, the evolution of the population direction selectivity nicely fits the predictions made by the total-income model where the income increases quasi-linearly with subsequent rewards (Fig. 5B, second panel). Our results together with the limited electrophysiological data reported in literature suggest that the total income might be coded in the directional selectivity strength of prefrontal cortex neurons, where reward-selective neurons produce the linear increment in selectivity. Further work is needed to quantify the correlation between the selectively strength and the total income of the matching behavior model.

To conclude, we put forward two hypotheses about the neural representations of the two behavioral models we considered here. First, we suggest that the three classes of neurons found using the first approach mentioned in the introduction (Wise and Murray, 2000; Brasted and Wise, 2005; Suzuki and Brown, 2005) actually code for the creation and dissolution of response–outcome associations and for error-prediction signals. In other words, the monotonic increases and decreases in neural firing rate correlating with the probability of correct response of the first two classes of neurons are brought about by the third population of neurons coding for error-prediction signals. Secondly, we suggest that the linear increase in direction selectivity found in Miller's lab (Asaad et al., 1998; Pasupathy and Miller, 2005) is a neural correlate of the total

income gained by the animal during learning, whose increase is produced by a population of neurons coding for the type of reward. The computations predicted by the two models could be implemented at different levels in the fronto-striatal loop: at the single neuron level (associative strengths) and the population level (total income).

## Acknowledgement

Andrea Brovelli was supported by a two-year post-doc fellowship awarded by the *Fondation pour la Recherche Médicale* (Paris, France).

## References

Asaad, W., Rainer, G., 1998. Neural activity in the primate prefrontal cortex during associative learning. Neuron 21, 1399–1407.

Balleine, B., Dickinson, A., 1998. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37, 407–419.

Brasted, P., Wise, S., 2004. Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. Eur. J. Neurosci. 19, 721–740.

Brasted, P., Wise, S., 2005. The arbitrary mapping of sensory inputs to volontary and involontary movements: Learning-dependent activity in the motor cortex and other telencephalic networks.

Cahusac, P., Rolls, E., Miyashita, Y., Niki, H., 1993. Modification of the responses of hippocampal neurons in the monkey during the learning of a conditional spatial response task. Hippocampus 19, 29–42.

Chen, L., Wise, S., 1995a. Neuronal activity in the supplementary eye field during acquisition of conditional oculomotor associations. J. Neurophysiol. 73, 1101–1121.

Chen, L., Wise, S., 1995b. Supplementary eye field contrasted with the frontal eye field during acquisition of conditional oculomotor associations. J. Neurophysiol. 73, 1122–1134.

Coquelin, P.A., Deguest, R., Munos, R., 2007. Sensitivity analysis in feynman-kac models. Tech. Rep.

Dickinson, A., 1997. Contemporary Animal Learning Theory. Cambridge University Press, UK.

Dickinson, A., 1994. Instrumental Conditioning. In: Mackintosh, N.J. (Ed.), Animal cognition and learning. Academic Press, London, pp. 45–79.

Doucet, A., Freitas, N.D., Gordon, N., 2001. Sequential Monte Carlo Methods in Practice. Springer.

Doucet, A., Godsill, S., 2000. On sequential monte carlo sampling methods for bayesian filtering. Stat. Comput. 10, 197–208.

Hadj-Bouziane, F., Boussaoud, D., 2003. Neuronal activity in the monkey striatum during conditional visuomotor learning. Exp. Brain Res. 153 (2), 190–196.

Herrnstein, R., 1970. On the law of effect. J. Exp. Anal. Behav. 13, 243–266.

Herrnstein, R.J., 1961. Relative and absolute strength of response as a function of frequency of reinforcement. J. Exp. Anal. Behav. 4, 267–272.

Macintosh, N., 1975. A theory of attention: variations in the associability of stimuli with reinforcement. Psychol. Rev. 82, 276–298.

Myung, I.J., 2003. Tutorial on maximum likelihood estimation. J. Math. Psychol. 47, 90–100.

Pasupathy, A., Miller, E., 2005. Different time courses of learning-related activity in the prefrontal cortex and striatum. Nat. 433 (7028), 873–876.

Pearce, J., Bouton, M., 2001. Theories of associative learning in animals. Annu. Rev. Psychol. 52, 111–139.

Pearce, J., Hall, G., 1980. A model for pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. 87, 532–552.

Rescorla, R.A., 1991. Associative relations in instrumental learning: The 18th barlett memorial lecture.

Rescorla, R.A., Wagner, A.R., 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Theory and Research. Appleton-Century-Crofts, New York, pp. 64–99.

Samejima, K., Doya, K., Ueda, K., Kimura, M., 2004. Estimating internal variables and parameters of a learning agent by a particle filter. Adv. Neural Inform. Process. Syst. 16, 1335–1342.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. Annu. Rev. Psychol. 57, 87–115.

Schultz, W., Dickinson, A., 2000. Neural coding of prediction errors. Annu. Rev. Neurosci. 23, 473–500.

Seung, S., 2004. Operant matching. <http://hebb.mit.edu/courses/9.29/2004/lectures/matching1.pdf>.

Spall, J.C., 2003. Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley.

Sugrue, L., Corrado, G., Newsome, W., 2004. Matching behavior and the representation of value in the parietal cortex. Science 304 (5678), 1782–1787.

Sugrue, L., Corrado, G., Newsome, W., 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. Nat. Rev. Neurosci. 6 (5), 363–375.

Sutton, R., Barto, A., 1998. Reinforcement Learning: An Introduction. MIT Press.

Suzuki, W., Brown, E.N., 2005. Behavioral and neurophysiological analyses of dynamic learning processes. Behav. Cogn. Neurosci. Rev. 4, 67–95.

Tremblay, L., Hollerman, J., Schultz, W., 1998. Modifications of reward expectation-related neuronal activity during learning in primate striatum. J. Neurophysiol. 80 (2), 964–977.

Tremblay, L., Schultz, W., 2000. Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. J. Neurophysiol. 83 (4), 1877–1885.

Watkins, C., Dayan, P., 1992. Q-learning. Mach. Learn. 8, 279–292.

Williams, Z., Eskandar, E., 2006. Selective enhancement of associative learning by microstimulation of the anterior caudate. Nat. Neurosci. 9 (4), 562–568.

Wirth, S., Yanike, M., Frank, L., Smith, A., Brown, E., Suzuki, W., 2003. Single neurons in the monkey hippocampus and learning of new associations. Science 300 (5625), 1578–1581.

Wise, S., Murray, E., 2000. Arbitrary associations between antecedents and actions. Trends Neurosci. 23, 271–276.